Data Needs for Consumer and Retail Firm Studies

Jeffrey M. Perloff and Mark Denbaly

Growing concentration in the retail grocery sector raises new economic questions that are difficult to answer with existing data sources. The data problems are due in large part to concentration in the retail data industry, where data are collected for commercial rather than academic research. Currently available grocery-level datasets are extremely expensive, are not properly randomized, and lack critical information.

To focus our discussion, we address data needs for industrial organization and marketing, nutrition and food safety, and government policy studies. The growing concentration at the grocery retail level raises a variety of industrial organization and marketing questions, such as: Has this greater concentration increased market power or changed the vertical relationship between manufacturers and other suppliers with retailers? Has the entry of low-price superstores fundamentally changed the services provided, the degree of product differentiation, the provision of private label products, and other actions by traditional supermarkets? What caused the mergers to occur?

Similarly, we want to know if greater concentration has affected the nation's nutrition and food safety, such as by making catastrophic food safety disasters more likely. Have increased product differentiation and lower prices from changes in retailing contributed substantially to alarming increases in rates of obesity?

Finally, we want to know how government rules and regulations have affected these markets and consumers. To protect consumers' health, the government has imposed restrictions on selling certain goods when food safety issues arise (e.g., mad cow disease and E. coli in lettuce and spinach). The government also provides nutritional and other label information (e.g., concerning health foods and organic foods) to help consumers make more informed food choices. What have been the effects of these laws and regulations on markets and on the health of various groups of consumers? We discuss the increase in concentration at the retail level, commercial databases, data needs for a number of important research areas, and possible solutions.

Concentration in Retail Markets

Grocery retailing markets are much more concentrated today than they were two decades ago. This increased concentration has altered the relationship between manufacturers and retailers. Although most existing empirical studies based on grocery scanner data implicitly presume that manufacturers set prices and retailers passively add on a competitive markup, there is substantial evidence (e.g., Villas-Boas) that such a description of the market is no longer true, if it ever was.

Mergers and acquisitions by large grocery retailers, including Kroger Co., Albertson's, Ahold USA, and Safeway, have significantly increased concentration ratios. Between 1997 and 2000, more than 4,100 U.S. supermarkets were acquired, representing $69 billion in sales. The four-firm concentration ratio (C4) increased from 16.6 percent in 1992 to 35.5 percent in 2005 (see figure 1). This trend toward increased concentration has continued with Supervalu's acquisition of third-ranked Albertson's in 2006 and the growth of Wal-Mart (Kaufman 2007).

Jeffrey Perloff is Professor, Department of Agricultural and Resource Economics, and member of the Giannini Foundation, University of California at Berkeley. Mark Denbaly is Deputy Director for Data and Web Communication, Economic Research Service, U.S. Department of Agriculture.

The opinions expressed in this paper are those of the authors and not necessarily the United States Department of Agriculture or any of its members. We thank Rui Huang for statistical help.

This article was presented in a principal paper session at the AAEA annual meeting (Portland, OR, July 2007). The articles in these sessions are not subjected to the journal's standard refereeing process.

Amer. J. Agr. Econ. 89 (Number 5, 2007): 1282–1287

DOI: 10.1111/j.1467-8276.2007.01097.x

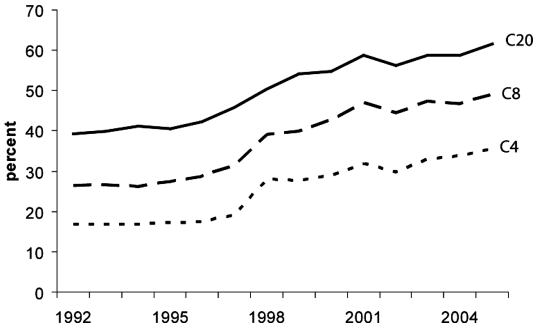

Figure 1. Top four (C4), eight (C8), and twenty (C20) firms' share of the U.S. grocery store sales

Companies that were not involved in the food business two decades ago, such as Wal-Mart and Target, now account for a significant share of consumers' food-at-home expenditures. Since 1994, nontraditional food retailers (supercenters, warehouse clubs, mass merchandisers, drugstores, and dollar stores) have steadily increased their market share by about 28 percentage points to 31.6 percent in 2005. Led by Wal-Mart, most of this growth is attributed to supercenters that commanded 17.1 percent of the food-at-home retail markets in 2005 (Kaufman 2007).

It took Wal-Mart just four years of aggressive supercenter growth to become the largest U.S. grocery chain by 2002. Wal-Mart's large share is due to its relatively low prices, which are driven by scale economies and efficient operations based on buying directly from suppliers. Wal-Mart's approach has started a domino effect, significantly changing the retail food market's landscape. Warehouse club and mass-merchandisers have adopted this strategy, further intensifying price competition as more consumers have switched from shopping at supermarkets to low-price, large-scale operations.

Many supermarkets and other traditional grocery retailers have reacted by expanding their operations through merger and acquisition strategies, introducing a wider variety of new products (e.g., organic and natural foods, upgrade store brands, and convenience foods), promoting new store formats, introducing self-checkout stations, expanding frequent shopper card programs, and offering online home shopping services. Some researchers contend mergers and acquisitions are driven by a search for efficiencies associated with consolidation as supermarkets are increasingly pressured to meet price competition from nontraditional food retailers like Wal-Mart. Others contend that mergers increase the market power of supermarkets and increase prices for consumers.

Growing retail concentration has not only changed the nature of competition at the retail level, it has greatly affected the vertical relations along the marketing chain. As a result of the competitive pressures from Wal-Mart and other nontraditional formats, many firms in the grocery industry have resorted to what the industry refers to as efficient consumer response. These methods are designed to enhance timely, accurate, continuous, consistent flow of products that are matched to consumer demands. The initiative focuses on reengineering activities in the selection of product assortments, product replenishment, product promotions, and new product introductions. Information on the type and extent of these business practices are not readily available, thus impeding efforts to examine their impact on prices and consumer welfare. Further, many researchers believe that the now larger retail vendors are exercising their increased oligopsony power to lower prices paid to suppliers and increasingly charging manufacturers slotting fees, which are lump-sum fees for carrying a new product or continuing to carry an existing one.

Commercial DataBases

Agricultural economists have studied a variety of demand, health, marketing, and industrial organization questions using data from grocery chains or proprietary retail grocery scanner data. Stores' loyalty card datasets do not include detailed information on household demographics and are potentially subject to more measurement errors due to infrequent use of loyalty cards or use of someone else's card for convenience. Moreover, grocery chains rarely make their databases available to researchers.

Today, the only two major firms providing such scanner data are Information Resources, Inc. (IRI) and Nielsen (formerly known as AC-Nielsen). Their datasets are constructed primarily for marketing purposes and are used by retailers, manufacturers, and farm commodity groups. Usually, these firms charge researchers prices comparable to those they charge their commercial customers, so that a dataset covering only a few commodities for the most recent year may cost hundreds of thousands of dollars.

The current major point-of-sale or store scanner data sources are IRI's InfoScan and Nielsen's ScanTrack. Store scanner data are collected at cash registers, while household scanner data are obtained from a sample of households that scan their purchases after each shopping trip. Over the past ten years, IRI and Nielsen also have begun to track grocery purchases by specific households. Nielsen's household scanner dataset is Homescan and IRI's is Consumer Network. (Knowledge Networks is also developing a household-based scanner data panel.)

These datasets provide richer household demographic information than are available in store scanner data (Muth, Siegel, and Zhen 2007). Because IRI and Nielsen instruct the household scanner data panelists to scan all purchases from all outlets, the datasets from household-based scanner data are more complete than grocery datasets of purchases of individual households collected through loyalty card users.

In addition to being expensive, commercial datasets come with significant restrictions on how they may be used (e.g., brand market shares may not be reported) and do not provide all critical information needed for many important research topics. For example, although feasible, they do not have information on whether a specific low-income household is a Women, Infants, and Children (WIC) program participant, they do not provide any details on retailers' cost of operation (e.g., wholesale prices), and the household scanner databases lack prices of nonpurchased items for demand studies.

Because scanner data are proprietary and are not primarily designed for academic research, detailed documentation on sampling and data collection procedures and statistical properties of the data are not readily available. Although few academic papers that use IRI and Nielsen data discuss the quality of these datasets, there is good reason to question whether these firms use proper random sampling techniques. In the store-based scanner data, large, traditional supermarket chains are over-represented (because they supply data and hence are included with certainty, as opposed to smaller stores that are sampled). In addition, store-based scanner data may not adequately include new sources of food sales (Wal-Mart supercenters and other big box stores, and WIC-only stores).

Muth, Siegel, and Zhen (2007) document the data collection process for Nielsen's Homescan data and identify potential sources of bias: sample design, self-selection, self-reporting, nonresponse, and attrition. However, no formal statistical studies have been conducted to measure the magnitude of the actual presence or the size of any potential bias. The households included in the sample are not probability based and randomly drawn from the community, and hence Homescan is a convenience sample.

We compared the U.S. Census demographic information with sample averages from IRI InfoScan by zip code area for all the zip coeds in the 1999 IRI dataset. Table 1 shows the averages across the zip code areas. IRI values could differ from Census data because only a subset of grocery stores is sampled within any given zip code or because the sampled households who shop at those grocery stores are not representative. In our sample, the IRI sample values have relatively large standard errors, so that we cannot conclude that the means of demographic variables in the Census and IRI datasets differ statistically significantly. However, in most zip codes areas, IRI households are younger, more likely to be white, larger, and more likely to be neither poor nor rich than are Census households (that is, typically large, white middle-class families).

Data Problems for Research

Purveyors of proprietary scanner data focus on the most recent marketing information for the industry and not on creating datasets that are ideal for research. In the proprietary datasets, short time series and lack of information from other levels of the production chain and other

Table 1. Comparison of U.S. Census and IRI Demographic Data

Households	IRI	Census
With individuals <18 years old	35.0%	33.9%
With income <$10,000	5.6%	7.4%
With income >$100,000	1.9%	14.3%
White	86.4%	71.9%
Black	5.3%	10.8%
Asian	1.3%	5.9%
Hispanic	5.7%	17.0%
Size	2.8	2.6

Notes: Average across all the zip code regions in the IRI data set. IRI data are for 1999 and Census data are from 2000.

missing variables limit the type of academic studies that are possible.

Industrial Organization and Marketing Studies

These datasets lack information that would facilitate studies of market power and vertical relations between manufacturers and retailers or suppliers.

To study markups over the food chain, other vertical relations, and food safety questions, we need to trace goods from the farm to the consumer. Most industrial organization studies and many nutritional and other studies require one to estimate a system of demand equations, which is often difficult with existing databases for three reasons.

First, the relevant prices are usually unavailable. Household datasets include prices for only purchased goods. In a few cases, researchers have matched store-level data with household data (or purchases by other households) to obtain the missing prices. Disturbingly, the price data from the grocery dataset do not always match that from the household dataset, and we lack any means of reconciling these differences.

Second, the actual transaction price is not obvious from the reported information. It is not possible to determine if the price reflects all discounts, coupons, and taxes. The commercial databases do not record whether the purchases were made using food stamps or WIC vouchers, which preclude studies of such programs and may bias standard demand equation estimates.

Third, the databases do not report shelf space allocations, local restrictions, or store warnings, all relevant advertising, information provided on the products (e.g., fat, health, safety, price per unit, and whether the product is organic), wholesale prices, slotting allowances, other transfers and restrictions between manufacturers and retailers, and government program information (e.g., WIC and food stamps).

Because the databases cover only a nonrandom subset of stores, conducting industrial organization studies of horizontal competition between stores is difficult. We do not have a complete enough set of stores to conduct spatial studies of pricing and other subjects. Research findings on the economics of consumer behavior provide insights into the effects of neighborhood characteristics on consumers' choices in differentiated product markets (cf. Waldfogel 2003).

Nutritional and Food Safety Studies

The high societal costs associated with obesity have intensified the need to identify and understand the factors that influence food choices and the effects of these choices on an individual's health. Extensive studies on consumer food demand show that food choices may depend on food prices, income levels, time available to shop and prepare meals, human capital resources, such as education and type of employment, and consumers' attitude, perception, diet, and nutrition knowledge, as well as psychological factors. Economic studies of these issues are greatly hampered by a lack of consistent and integrated data and information.

No single reliable source currently provides or could provide all of the information required for a myriad of studies that could be undertaken. A number of data sources do provide some of the information, but each is weak in critical areas. A 2005 report by the National Research Council of the National Academies (NRC) recommended enhancing usability of various key data systems to support research on critical U.S. food and nutrition policies. Adopting the NRC's recommendation to create integrated and consistent data would help researchers to better understand how consumers' food choices, diets, and health are affected by changes in food prices, neighborhood characteristics, access to food stores and restaurants, behavioral factors, and by participation in food assistance programs.

The National Health and Nutrition Examination Survey (NHANES), conducted by the National Center for Health Statistics of the Centers for Disease Control, measures food intakes and an array of health outcomes for a representative population, but no information on prices of foods eaten by survey respondents is collected. Adding price information from other existing sources would enable research on drivers of consumer food choice and their connections to health outcomes for various population subgroups and regions overtime. Measuring consumer price responsiveness is a critical component of a sound policy strategy. Beyond characterizing consumer preferences, information on price responsiveness enables researchers to evaluate the effects of taxes and subsidies on consumption of various foods

and the nutrients they contain. Further, without controlling for price variations, researchers cannot consistently estimate the role of other factors. Adding data and information on consumer attitude, perception, diet, and nutrition knowledge, and psychological factors to the NHANES intake data would facilitate studies of the drivers of the obesity epidemic.

Currently, no dataset provides the capability to trace foods back to their sources. Plans of Wal-Mart and others to use radio signals (RFID tags) to track goods from the manufacturer to the retailer or final consumer raise privacy concerns but also may provide a means to examine important questions concerning food safety, food quality, and various vertical integration issues. However, we know of no plans to make such information available to researchers. Indeed, manufacturing and retailing firms may not want such information disseminated.

Nutritional studies are hampered by a lack of datasets that cover both food at home and food in restaurants. As Americans have increasingly switched from home-cooked meals to processed foods or restaurant meals, the substitution patterns between these types of meals has substantial public policy importance.

Government Programs

Many studies of government programs require time series data. Bizarrely, both IRI and Nielsen usually discard data that are more than three years old, making many time series or historical studies of government laws and regulations difficult or impossible to conduct. For example, data from these sources before and after the recent change in the U.S. rules on organic foods are generally not available either because datasets.

Food assistance programs are designed to provide a nutritional safety net, guaranteeing a minimum level of access to essential nutrients for participants. Empirical evidence on the extent to which the programs affect consumption, nutrient intake, and obesity provides critical information about the current effectiveness of the programs. Combining the existing measures of consumption patterns and the health status of program participants with this information on benefit levels and duration of participation will help to reveal the critical link between food assistance programs and the diet, nutrition, and health outcomes of program participants. For example, accounting for how long participants have been in the sample can help researchers determine if the sizes of the program's effects differ depending on the duration of participation.

The NHANES queries respondents about their program participation and benefits. However, studies show that self-reported information is systematically underreported in many surveys, including NHANES. For example, in 2004, the Current Population Survey captured 60% of average monthly caseloads and 58% of annual benefits (Bollinger and David 2005). Administrative records can be used to correct this underreporting and avoid analytical results that would otherwise be biased.

Supplementing the NHANES dataset with administrative records would allow researchers to study the connection between food choices and neighborhood characteristics, particularly for low-income households in urban and rural areas. To the extent that NHANES includes such households, researchers could correlate health and nutrition outcomes with household and location characteristics. A link between NHANES data and information on the location of food stores and eating establishments would also enhance efforts to understand the effects of access on food choices and health outcomes. Information on locations and characteristics of food stores and foodservice establishments can be collected using proprietary sources, such as Spectra® and NPD. Linking NHANES to household and local community descriptors in the Census's American Community Survey will help researchers understand how neighborhood characteristics influence food choices and health outcomes.

Improving Datasets

We have a simple and obvious message. With more data, economists could analyze additional, important issues of economic theory and government policies.

Because data lack rivalry (everyone can consume the data), society under-provides data. Relying on commercial vendors is unattractive because these firms charge very high prices, do not fully disclose the nature of their data, provide data for only very short periods, and report only variables that are important for commercial customers and not all variables that are important for researchers.

One approach to ameliorating data shortages for research would be to have government agencies or nonprofit organizations collect the ideal datasets or provide incentives to commercial providers. Fundamentally, researchers need access to unrestricted data based on proper random samples and that include all the relevant variables.

First, to enable unfettered assess, to improve content, and to obtain better prices, it may make sense for university and government researchers and organizations (the AAEA, government agencies, business school organizations, the American Economic Association, and others) to try to negotiate with private purveyors collectively. They might also negotiate to house, at little or no cost, historical IRI and Nielsen data that are now discarded so that longer time series and additional variables can be created. However, such collective action might raise antitrust issues.

Second, these research groups could try to make arrangements with individual firms to supply data. We know of at least two supermarket chains that have been willing to make such agreements in the past. The AAEA could lead efforts to select representative samples of suppliers to collect details of proprietary transaction data and provide them to researchers so that privacy and confidentiality of the data are maintained.

Third, these research organizations could collaborate to collect data on their own. Even discussing this possibility may facilitate negotiations with commercial data purveyors.

On a less grand scale, we have a laundry list of new datasets that would be particularly useful. First, industrial organization and food safety studies require information at both the retail and upstream levels, including information about wholesale prices, food sources, various slotting and tying relations, and government programs.

Second, nutritional studies need datasets that combine information on food-at-home and away-from-home, nutritional content of these various foods, and prices. Because consumer studies find substantial variation in nutritional consumption across demographic groups and neighborhoods, datasets are needed that cover a broad cross-section.

Third, health and nutrition studies would benefit substantially if we could link the intake and health data with administrative food assistance records to add levels and duration of program assistance. Such a link would have to address two challenging issues: (1) privacy and confidentiality conditions under which states collect the administrative data must be met to access the data for linking purposes and (2) variation of data formats across states makes linking these sets to survey data difficult. In addition, given the relatively small effects of price and income on food choices, addressing the obesity epidemic may require collection of new data on consumers' health and nutritional knowledge, attitudes, and available time to shop and prepare meals to undertake economic studies to understand consumer dietary behavior.

References

Bollinger, C., and M. David. 2005. "I Didn't Tell, and I Won't Tell: Dynamic Response Error in the SIPP." *Journal of Applied Econometrics* 20:563–69.

Kaufman, P. 2007. "Food Market Structures: Food Retailing." *U.S. Department of Agriculture, Economic Research Service*. www.ers.usda.gov/Briefing/FoodMarketStructures/foodretailing.htm (2007).

Muth, M.K., P.H. Siegel, and C. Zhen. 2007. "ERS Data Quality Study Design." Final Report, Research Triangle Institute, Project Number 210153.001.

National Research Council of National Academies. 2005. "Improving Data to Analyze Food and Nutrition Policies." Committee on National Statistics, Panel on Enhancing the Data Infrastructure in Support of Food and Nutrition Programs, Research, and Decision Making, National Academies Press, Washington, DC.

Villas-Boas, S.B. 2007. "Vertical Relationships Between Manufacturers and Retailers: Inference With Limited Data." *Review of Economic Studies* 74:625–52.

Waldfogel, J. 2003. "Preference Externalities: An Empirical Study of Who Benefits Whom in Differentiated Product Markets." *Rand Journal of Economics* 34:557–68.